A general convergence method for Reinforcement Learning in the continuous case

Rémi Munos

CEMAGREF, LISC, Parc de Tourvoie
BP 121, 92185 Antony Cedex, FRANCE
E-mail: Remi.Munos@cemagref.fr

Abstract. In this paper, we propose a general method for designing convergent Reinforcement Learning algorithms in the case of continuous state-space and time variables. The method is based on the discretization of the continuous process by convergent approximation schemes : the Hamilton-Jacobi-Bellman equation is replaced by a Dynamic Programming (DP) equation for some Markovian Decision Process (MDP).
If the data of the MDP were known, we could compute the value of the DP equation by using some DP updating rules. However, in the Reinforcement Learning (RL) approach, the state dynamics as well as the reinforcement functions are a priori unknown, leading impossible to use DP rules.
Here we prove a general convergence theorem which states that if the values updated by some RL algorithm are close enough (in the sense that they satisfy a "weak" contraction property) to those of the DP, then they converge to the value function of the continuous process. The method is very general and is illustrated with a model-based algorithm built from a finite-difference approximation scheme.

1 Introduction

This paper proposes a convergence result for Reinforcement Learning (RL) algorithms in the case of continuous state-space and time variables. RL uses the method of Dynamic Programming (DP) which defines the optimal feed-back control by approximating the *value function*, which is the best future cumulative reinforcement as a function of initial state.

A classical approach in optimal control for computing the value function consists in using approximation schemes (deduced from finite-element or finite-difference methods) which replace the continuous process by a discrete one (see [KD92]) for some given resolution. We obtain a finite Markovian Decision Process (MDP) whose DP equation may be computed by classical value iteration DP rules, knowing that (in the discounted case) the convergence of this method is guaranteed by some "strong" contraction property satisfied by the updated values.

However, in the RL approach, the state dynamics as well as the reinforcement functions are considered (at least partially) unknown from the system. Thus the values of the DP updating rule are unknown and the "strong" contraction property is no more valid.

This paper states that if this contraction property is weakened, we still have the convergence of the method as the resolution of the discretization tends to zero and the number of iterations tends to infinity. *This result allows approximation while keeping the convergence.* The theorem is very general and may apply for a wide class of RL algorithms such as model-based or model-free algorithms, with some "on-line" or "off-line" updating rule, for deterministic or stochastic state dynamics. We propose an example of model-based algorithm in the deterministic case whose values satisfy the "weak" contraction property, thus insuring its convergence.

Section 2 proposes the formalism for optimal control problems in the continuous case. The method of DP is described : the value function is introduced and the Hamilton-Jacobi-Bellman (HJB) equation is stated. A finite-difference approximation scheme is detailed and a theorem of convergence for the scheme is stated. *Section 3* is concerned with RL algorithms. The general theorem is stated and its proof is given. Then an example of model-based algorithm built from a finite-difference scheme is described and the proof that the computed values satisfy the "weak" contraction property is given in *appendix A*.

2 The Optimal Control Formalism

We illustrate our method in the particular case of *deterministic* controlled systems with *infinite time horizon* and *discounted reinforcement*. A study of the stochastic case may be found in [MB97].

Let $x(t) \in \bar{O}$ be the state of the system with O an open and bounded subset of $\mathbb{R}^d$. The evolution of the system (its *state dynamics* f) depends on the *current state* $x(t)$ and *control* $u(t)$; it is defined by a controlled differential equation :

$$\frac{d}{dt}x(t) = f(x(t), u(t)) \tag{1}$$

where the control $u(t)$ is a bounded, Lebesgue measurable function with values in a compact U. From any initial state x, the choice of a control $u(t)$ leads to a unique *trajectory* $x(t)$. Let τ be the *exit time* of $x(t)$ from $\bar{O}$ (with the convention that if $x(t)$ always stays in $\bar{O}$, then $\tau = \infty$). Then, we define the discounted reinforcement functional of state x, control $u(.)$:

$$J(x; u(.)) = \int_0^\tau \gamma^t r(x(t), u(t))dt + \gamma^\tau R(x(\tau))$$

Where $r(x, u)$ is the *running reinforcement* and $R(x)$ the *boundary reinforcement*. γ is the *discount factor* $(0 \leq \gamma < 1)$.

The **objective of the control problem** is to find the optimal control (which can be expressed here as a feed-back law $u^*(x)$) that optimizes the reinforcement functional for any state x.

2.1 The Method of Dynamic Programming (DP)

The DP method computes the optimal control by introducing the *value function*, maximal value of the functional as a function of initial state x :

$$V(x) = \sup_{u(.)} J(x; u(.)) \tag{2}$$

Following the DP principle, we prove that the value function satisfies a first-order nonlinear partial differential equation called the *Hamilton-Jacobi-Bellman* equation (see [FS93] for a survey) (in the stochastic case, it is of a second order).

Theorem 1 : Hamilton-Jacobi-Bellman. *If V is differentiable at $x \in O$, let $DV(x)$ be the gradient of V at x, then the following HJB equation holds at x.*

$$V(x) \ln \gamma + \sup_{u \in U} [DV(x).f(x,u) + r(x,u)] = 0$$

Hypotheses 1 In the following, we assume that :
- f and r are bounded with M_f (respectively M_r) and Lipschitzian : $|f(x,u) - f(y,u)| \leq L_f \|x - y\|_1$ (resp. $|r(x,u) - r(y,u)| \leq L_r \|x - y\|_1$), with the norm $||x||_1 = \sum_{i=1}^{d} |x_i|$.
- R is Lipschitzian : $|R(x) - R(y)| \leq L_R \|x - y\|_1$.
- The boundary ∂O is $\mathcal{C}^2$.

Besides, we consider the following hypothesis concerning the state dynamics around the boundary, and we state a result of continuity for V (see [Bar94]).

Hypothesis 2 For all $x \in \partial O$, let $\overrightarrow{n}(x)$ be the outward normal of O at x, we assume that :
- *If* $\exists u \in U$, s.t. $f(x,u).\overrightarrow{n}(x) \leq 0$ *then* $\exists v \in U$, s.t. $f(x,v)\overrightarrow{n}(x) < 0$.
- *If* $\exists u \in U$, s.t. $f(x,u).\overrightarrow{n}(x) \geq 0$ *then* $\exists\, v \in U$, s.t. $f(x,v)\overrightarrow{n}(x) > 0$.

Theorem 2 : Continuity. *Suppose that these hypotheses hold, then the value function is continuous in O.*

2.2 Approximation Schemes

In order to approximate the value function, we use the numerical schemes (for example based on finite-element (FE) or finite-differences (FD) methods) of Kushner [KD92], which replace the continuous problem by a discrete one. The HJB equation is discretized, for some resolution δ, by a DP equation for some MDP, whose value is V^δ. We state that the value V^δ of the approximation scheme converges to the value function V of the continuous process as the discretisation step δ tends to 0. As an illustration, we describe here the FD method.

Description of a FD scheme : Let $e_1, e_2, ..., e_d$ be a basis for $\mathbb{R}^d$. The state dynamics is : $f = (f_1, ..., f_d)$. Let the positive and negative parts of f_i be : $f_i^+ = \max(f_i, 0)$, $f_i^- = \max(-f_i, 0)$.

For any discretization step δ, we consider the lattice $\delta\mathbb{Z}^d = \left\{\delta . \sum_{i=1}^d j_i e_i\right\}$ where $j_1, ..., j_d$ are any integers, and define :

- the **discretized state space :** $\Sigma^\delta = \delta\mathbb{Z}^d \cap O$ (see figure 1), and

- its **frontier** $\partial\Sigma^\delta = \{\xi \in \delta\mathbb{Z}^d \setminus \Sigma^\delta$, such that at least one adjacent points $\xi \pm \delta e_i \in O\}$

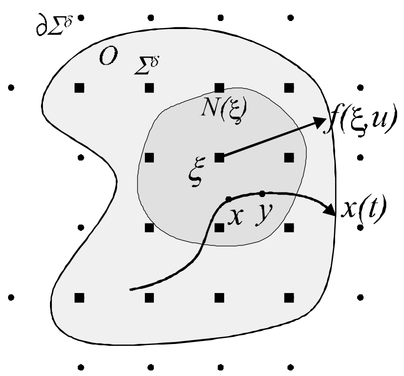

Fig. 1. The discretized state space Σ^δ (the square dots) and its frontier $\partial\Sigma^\delta$ (the round dots). A trajectory $x(t)$ crosses the neighbourhood $N(\xi)$ (in dark grey) of vertex ξ. Let 2 points of the trajectory $x = x(t_0)$ and $y = x(t_0 + \tau)$ be such that the control u is kept constant during $t \in [t_0, t_0 + \tau]$. Then we make the model $\widetilde{f}(\xi, u) = \frac{y-x}{\tau}$ of the state dynamics $f(\xi, u)$.

We approximate the control space U by some finite control spaces $U^\delta \subset U$ such that for $\delta \leq \delta'$ we have $U^{\delta'} \subset U^\delta$ and besides, $\overline{\cup_\delta U^\delta} = U$.

The FD approximation consists in discretizing the HJB equation by :

$$V^\delta(\xi)\ln\gamma + \sup_{u\in U^\delta}\left\{\sum_{i=1}^d \left[f_i^+(\xi,u).\Delta_i^+ V^\delta(\xi) + f_i^-(\xi,u).\Delta_i^- V^\delta(\xi)\right] + r(\xi,u)\right\} = 0 \tag{3}$$

where the gradient $DV(\xi)$ is replaced by the forward and backward difference quotients of V at ξ in direction $i = 1..d$:

$$\Delta_i^+ V(\xi) = \frac{1}{\delta}\left[V(\xi + \delta e_i) - V(\xi)\right]$$
$$\Delta_i^- V(\xi) = \frac{1}{\delta}\left[V(\xi - \delta e_i) - V(\xi)\right]$$

Knowing that $(\Delta t \ln \gamma)$ is an approximation of $(\gamma^{\Delta t} - 1)$ as Δt tends to 0, we obtain from (3) the following equivalent equation: for $\xi \in \Sigma^\delta$,

$$V^\delta(\xi) = \sup_{u \in U^\delta} \left\{ \gamma^{\tau(\xi,u)} \sum_{\xi'} p(\xi, \xi', u).V^\delta(\xi_i) + \tau(\xi,u) r(\xi,u) \right\} \tag{4}$$

$$\text{with} : \tau(\xi,u) = \frac{\delta}{||f(\xi,u)||_1} \text{ and} : p(\xi,\xi',u) = \frac{f_i^\pm(\xi,u)}{||f(\xi,u)||_1} \text{ for } \xi' = \xi \pm \delta e_i \tag{5}$$
$$= 0 \text{ otherwise.}$$

This equation is a DP equation for a finite MDP (see [FS93]) whose *state space* is $\Sigma^\delta \cup \partial\Sigma^\delta$. Its *control space* is U^δ and the *probabilities of transition* $p(\xi,\xi',u)$ from the state ξ, to the next state ξ' with some control u are the normalized coordinates $\frac{|f_i(\xi,u)|}{||f(\xi,u)||_1}$.

Besides, we have the boundary condition $V^\delta(\xi) = R(\xi)$ for $\xi \in \partial\Sigma^\delta$.

Resolution of the scheme : By defining the approximation scheme F^δ, operator on the space of functions on Σ^δ :

$$F^\delta[W](\xi) = \sup_{u \in U^\delta} \left\{ \gamma^{\tau(\xi,u)} \sum_{\xi'} p(\xi,\xi',u).W(\xi_i) + \tau(\xi,u) r(\xi,u) \right\}, \tag{6}$$

equation (4) becomes $V^\delta = F^\delta\left[V^\delta\right]$. The solution V^δ may be computed by some DP value iteration method where V^δ is obtained as a limit of successive iterations:

$$V^\delta_{n+1} \leftarrow F^\delta\left[V^\delta_n\right]. \tag{7}$$

For any initial V^δ_0, we compute $V^\delta_1 \leftarrow F^\delta\left[V^\delta_0\right]$, then $V^\delta_2 \leftarrow F^\delta\left[V^\delta_1\right]$, and so on. Thank to the discounted factor γ, such updated values satisfy the following **"strong" contraction property** (with some $\lambda = 1 - \frac{\delta}{2M_f}\ln\frac{1}{\gamma}$):

$$\left\|V^\delta_{n+1} - V^\delta\right\|_{\Sigma^\delta \cup \partial\Sigma^\delta} \leq \lambda. \left\|V^\delta_n - V^\delta\right\|_{\Sigma^\delta \cup \partial\Sigma^\delta} \tag{8}$$

(with $||.||_\Xi$ denoting $\sup_{\xi \in \Xi}|.|$), from which we deduce that for any given discretisation step δ, the constant $\lambda < 1$ thus the values V^δ_n converge to V^δ as n tends to infinity.

Convergence of the scheme The following theorem, whose proof uses the general convergence result of Barles (see [BS91] and [Bar94]) and the strong comparison result between sub- and super- viscosity solution (see [FS93]) of HJB equations, insures that V^δ is a convergent approximation of V.

Theorem 3 : Convergence of the scheme. *Let us assume that the hypotheses 1 and 2 hold, then V^δ converges to V as δ tends to 0 :*

$$\lim_{\substack{\delta \downarrow 0 \\ \xi \to x}} V^\delta(\xi) = V(x) \textit{ uniformly on any compact } \Omega \subset O$$

Figure 2 summerizes the two previous results of convergence, which are : for any distretization step δ, the values V_n^δ computed by the DP updating rule (7) tend to the value V^δ of the DP equation (4) as n tends to infinity, and from the convergence of the scheme (theorem 3), V^δ tends to the value function V of the continuous process as δ tends to zero.

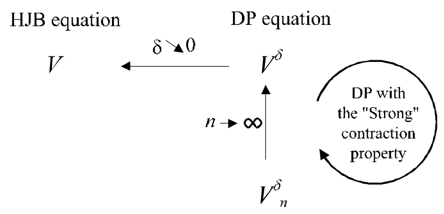

Fig. 2. The HJB equation is discretized, for some resolution δ, into a DP equation whose solution is V^δ. The convergence of the scheme insures that $V^\delta \longrightarrow V$ as $V \longrightarrow 0$. Thanks to the "strong" contraction property, the iterated values V_n^δ tend to V^δas $n \longrightarrow \infty$.

3 Reinforcement Learning

RL is a constructive and iterative process, based on experience, that intends to estimate the value function by successive approximations. Thus, **in the RL approach, we have the constraint that the state dynamics f and the reinforcement functions r, R are a priori unknown from the system**.

Thus the probabilities of transition $p(\xi, \xi', u)$ and the time $\tau(\xi, u)$ are unknown and have to be approximated. We deduce that the strong contraction property (8) cannot hold any more. However, we prove that if some weaker contraction property does hold, then we can obtain the convergence as well. The following section states the general convergence theorem for RL algorithms provided that the updated values satisfy some **"weak" contraction property**. *The statement of such "good approximations" satisfying this property is the basis for designing convergent algorithms.*

3.1 A general Theorem of Convergence

Theorem 4 : Convergence of RL algorithms. *Suppose that the values V_n^δ updated with some algorithm satisfy the following* **"weak" contraction property** *with respect to a solution V^δ of a convergent approximation scheme (such as (6)):*

$$\left\|V_{n+1}^\delta - V^\delta\right\|_{\Sigma^\delta} \leq (1 - k_1.\delta).\left\|V_n^\delta - V^\delta\right\|_{\Sigma^\delta} + e(\delta)\delta \tag{9}$$

$$\left\|V_{n+1}^\delta - V^\delta\right\|_{\partial\Sigma^\delta} \leq k_2.\delta \tag{10}$$

for some positive constants k_1, k_2 and some fonction $e(\delta) \searrow 0$ as $\delta \searrow 0$. Then for all $\varepsilon > 0$, there exists Δ and N, such that $\forall \delta \leq \Delta, \forall n \geq N$,

$$\sup_{\Sigma^\delta \cap \Omega} \left\| V_n^\delta - V \right\| \leq \varepsilon \text{ on any compact } \Omega \subset O.$$

Remark. Here, one cannot expect any more that for a given δ the values V_n^δ converge to V^δ. However the theorem states that the convergence occurs as $\delta \searrow 0$ and $n \to \infty$. Figure 3 summerizes this result.

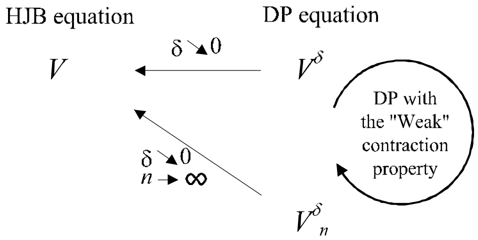

Fig. 3. When the "strong" contraction property property does not hold any more, one cannot expect that the computed values V_n^δ tend to V^δ. However, the theorem states that, thanks to the "weak" contraction property, the values V_n^δ tend to the value function V as $n \to \infty$ and $\delta \searrow 0$.

Proof of theorem 4. Let us denote $E_n^\delta = \left\| V_n^\delta - V^\delta \right\|_{\Sigma^\delta \cup \partial \Sigma^\delta}$. Let $\Omega \subset O$ be any compact. For any $\varepsilon > 0$, let us choose $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $\varepsilon_1 + \varepsilon_2 = \varepsilon$. From the convergence of the scheme, theorem 3 states that there exists Δ_1 such that for all $\delta \leq \Delta_1, \sup_{x\in\Omega} |V^\delta(x) - V(x)| \leq \varepsilon_1$. The idea is to prove that there exists Δ_2, for $\delta \leq \Delta_2$, there exists N, for all $n \geq N$,

$$E_n^\delta \leq \varepsilon_2. \tag{11}$$

Then we will obtain that for any $\delta \leq \Delta = \min\{\Delta_1, \Delta_2\}$, for all $n \geq N$,

$$\begin{aligned} \sup_{\xi \in \Omega \cap \Sigma^\delta} |V_n^\delta(\xi) - V(\xi)| &\leq \sup_{x \in \Omega} |V^\delta(x) - V(x)| + \sup_{\xi \in \Sigma^\delta \cup \partial \Sigma^\delta} |V_n^\delta(\xi) - V^\delta(\xi)| \\ &\leq \varepsilon_1 + \varepsilon_2 = \varepsilon \end{aligned}$$

A sufficient condition for (11) : Suppose that there exists a positive constant α such that the following conditions hold true :

$$\text{If } E_n^\delta > \varepsilon_2 \text{ then } \left\| V_{n+1}^\delta - V^\delta \right\|_{\Sigma^\delta} \leq E_n^\delta - \alpha \tag{12}$$

$$\text{If } E_n^\delta \leq \varepsilon_2 \text{ then } \left\| V_{n+1}^\delta - V^\delta \right\|_{\Sigma^\delta} \leq \varepsilon_2 \tag{13}$$

then we deduce that there exits N such that for $n \geq N$, $\left\|V_{n+1}^{\delta} - V^{\delta}\right\|_{\Sigma^{\delta}} \leq \varepsilon_2$. Besides, from the property (10), for $\delta \leq \frac{\varepsilon_2}{k_2}$, $\left\|V_{n+1}^{\delta} - V^{\delta}\right\|_{\partial\Sigma^{\delta}} \leq \varepsilon_2$, thus $E_n^{\delta} \leq \varepsilon_2$.

Proof of the sufficient condition:

Let us prove that for all $\varepsilon_2 > 0$, there exists Δ_2 such that for all $\delta \leq \Delta_2$, conditions (12) and (13) are satisfied. For any $\varepsilon_2 > 0$, from the convergence of $e(\delta)$ to 0 as $\delta \downarrow 0$, there exists Δ_2 such that for $\delta \leq \Delta_2$ the following condition hold:

$$e(\delta) - k_1.\frac{\varepsilon_2}{2} \leq 0 \tag{14}$$

First, suppose that $E_n^{\delta} > \varepsilon_2$, then from (9),

$$\left\|V_{n+1}^{\delta} - V^{\delta}\right\|_{\Sigma^{\delta}} \leq (1 - k_1.\delta)E_n^{\delta} + e(\delta).\delta \leq E_n^{\delta} - k_1.\delta.\varepsilon_2 + e(\delta).\delta.$$

and from (14), $\left\|V_{n+1}^{\delta} - V^{\delta}\right\|_{\Sigma^{\delta}} \leq E_n^{\delta} - k_1.\delta.\frac{\varepsilon_2}{2} + e(\delta).\delta - k_1.\delta.\frac{\varepsilon_2}{2} \leq E_n^{\delta} - k_1.\delta.\frac{\varepsilon_2}{2}$. Thus condition (12) hold true for $\alpha = k_1.\delta.\frac{\varepsilon_2}{2}$.

Now suppose that $E_n^{\delta} \leq \varepsilon_2$, then from (9),

$$\left\|V_{n+1}^{\delta} - V^{\delta}\right\|_{\Sigma^{\delta}} \leq (1 - k_1.\delta)\frac{\varepsilon_2}{2} + \frac{\varepsilon_2}{2} + e(\delta)\delta - k_1.\delta\frac{\varepsilon_2}{2} \leq \frac{\varepsilon_2}{2} + \frac{\varepsilon_2}{2} = \varepsilon_2$$

and condition (13) is true. Thus conditions (12) and (13) are true and for $\delta \leq \Delta = \min\{\Delta_1, \Delta_2, \frac{\varepsilon_2}{k_2}\}$, for all $n \geq N$, we have:

$$\begin{aligned}\sup_{\xi \in \Omega \cap \Sigma^{\delta}} |V_n^{\delta}(\xi) - V(\xi)| &\leq \sup_{x \in \Omega} |V^{\delta}(x) - V(x)| + \sup_{\xi \in \Sigma^{\delta} \cup \partial\Sigma^{\delta}} |V_n^{\delta}(\xi) - V^{\delta}(\xi)| \\ &\leq \varepsilon_1 + \varepsilon_2 = \varepsilon \quad \blacksquare\end{aligned}$$

This theorem provides a general method for designing convergent RL algorithms. It may apply to model-free (see [Mun96] or [Mun97]) or model-based algorithms, with on-line or off-line (for example synchronous, Gauss-Seidel, asynchronous) DP updating methods, and for deterministic or stochastic dynamics.

3.2 An Example of Model-Based Algorithm

The idea is to build a model of the state dynamics f and of the reinforcement function r at the vertices ξ of the discretization from samples of trajectories going through their neighbourhood. Then, from this model, we define the approximated transition probabilities which are used, instead of the exact ones $p(\xi, \xi', u)$, in the updating rule (7).

In the following, we assume that the state dynamics f is bounded from below (there exists m_f such that $\|f\|_1 \geq m_f$).

- **Estimation for** $\xi \in \Sigma^{\delta}$ **:** For any vertex $\xi \in \Sigma^{\delta}$, any control $u \in U^{\delta}$, we build a model $\widetilde{f}$ and $\widetilde{r}$, approximations of f and r from trajectories $x(t)$ going through the neighbourhood of ξ: we consider some states $x = x(t_0)$ and $y = x(t_0 + \tau)$ such that:
  - $x \in N(\xi)$ *neighbourhood* of ξ (whose diameter is inferior to $k_N.\delta$ for some positive constant k_N).

- the control u is kept constant for $t \in [t_0, t_0 + \tau]$,
- the time τ satisfies, for two positive constantes k_1 et k_2, the relation:

$$k_1 \delta \leq \tau \leq k_2 \delta. \tag{15}$$

See figure 1. Then we make the following model for state ξ and control u :

$$\widetilde{f}(\xi, u) = \frac{y - x}{\tau}$$
$$\widetilde{r}(\xi, u) = r(x, u)$$

Then we compute the approximated probabilities $\widetilde{p}(\xi, \xi', u)$ and time $\widetilde{\tau}(\xi, u)$ by using in the equations (5) the model $\widetilde{f}$ instead of f and obtain the following updating rule based on (7):

$$V_{n+1}^{\delta}(\xi) \leftarrow \gamma^{\widetilde{\tau}(\xi,u)}. \sum_{\xi'} \widetilde{p}(\xi, \xi', u).V_n^{\delta}(\xi') + \widetilde{\tau}(\xi, u).\widetilde{r}(\xi, u) \tag{16}$$

which can be used as an "off-line" (synchronous, Gauss-Seidel, asynchronous) or "on-line" (for example by updating $V_n^{\delta}(\xi)$ as soon as a trajectory leaves the neighbourhood of ξ) updating DP method (see [BBS95]).
- **Estimation for** $\xi \in \partial \Sigma^{\delta}$: As soon as a trajectory $x(t)$ exits from the state space at $y \in \partial O$, we consider the states $\xi \in \partial \Sigma^{\delta}$ whose respective neighbourhoods $N(\xi)$ contain y and we update their value with:

$$V_{n+1}^{\delta}(\xi) \leftarrow R(y) \tag{17}$$

The following theorem states that the algorithm consisting in updating regularly all the states $\xi \in \Sigma^{\delta}$ with rule (16) and all states $\xi \in \partial \Sigma^{\delta}$ (at least once each) with (17) satisfies the "weak" contraction property (9) and (10) thus defines a convergent algorithm. The proof is given in *appendix A*.

Theorem 5. Convergence of the model-based, FD algorithm. *The updating rules (16) and (17) satisfy the "weak" contraction property (9) and (10) with respect to the convergent approximation scheme (6), thus the theorem 4 applies and the model-based FD algorithm is convergent.*

4 Conclusion

We proposed a framework for designing RL algorithms and proving their convergence. The method is very general since the only required property is the "weak" contraction property with respect to some convergent approximation scheme. The choice of such a numerical scheme is free and may come from any discretization method such as finite difference or finite element method, using a constant or a variable resolution. As an illustration, we proposed a very simple model-based algorithm build from a finite-difference approximation scheme and proved its convergence.

A Convergence of the Model-Based FD Algorithm

A.1 Some Majorations

Comparison of the times $\tau(\xi, u)$ and $\widetilde{\tau}(\xi, u)$.

From the Lipschitz property of f, we have the following Taylor majoration:

$$||y - x - f(x,u).\tau||_1 \leq \tfrac{1}{2} L_f.\tau^2$$

Since the neighbourhood of ξ is of a diameter inferior to $k_N.\delta$, we have:

$$||f(x,u) - f(\xi,u)||_1 \leq L_f.||x - \xi||_1 \leq L_f k_N \delta.$$

But $||y - x - f(\xi,u).\tau||_1 = ||y - x - f(x,u).\tau + \tau[f(x,u) - f(\xi,u)]||_1$, thus from (15), we have: $||y - x - f(\xi,u).\tau||_1 \leq \left(\frac{k_2}{2} + k_N\right) L_f k_2 \delta^2$. And because the state dynamics f is bounded from below by m_f and that $\tau \geq k_1 \delta$, we have $||y - x||_1 \geq k_1 m_f \delta$, thus:

$$|\tau(\xi,u) - \widetilde{\tau}(\xi,u)| \leq k_\tau \delta^2 \tag{18}$$

with: $k_\tau = \frac{(\frac{k_2}{2} + k_N) L_f k_2}{k_1 m_f^2}$. We deduce, by using a property of the exponential function that:

$$\left|\gamma^{\tau(\xi,u)} - \gamma^{\widetilde{\tau}(\xi,u)}\right| \leq k_\tau \ln \frac{1}{\gamma}.\delta^2 \tag{19}$$

Comparison of the probabilities $p(\xi, \xi', v)$ and $\widetilde{p}(\xi, \xi', v)$.

For $\xi' \neq \xi \pm \delta e_i$, $\widetilde{p}(\xi,\xi',v) = p(\xi,\xi',v) = 0$ and for $\xi' = \xi \pm \delta e_i$, we have:

$$\left| \frac{\widetilde{f}_i^{\pm}(\xi,u)}{\left\|\widetilde{f}(\xi,u)\right\|_1} - \frac{f_i^{\pm}(\xi,u)}{||f(\xi,u)||_1} \right| \leq \frac{\left|\left\|\widetilde{f}(\xi,u)\right\|_1 - ||f(\xi,u)||_1\right| + \left|f_i^{\pm}(\xi,u) - \widetilde{f}_i^{\pm}(\xi,u)\right|}{||f(\xi,u)||_1}$$

From what precedes, $\left\|\widetilde{f}(\xi,u) - f(\xi,u)\right\|_1 \leq \left(\frac{k_2}{2} + k_N\right) L_f \delta$ and we deduce:

$$|\widetilde{p}(\xi,\xi',v) - p(\xi,\xi',v)| \leq \frac{L_f}{m_f}(k_2 + 2k_N)\delta \tag{20}$$

A.2 Convergence of the Model-Based FD Algorithm

The value $V_n^\delta(\xi)$ is updated with:

$$V_{n+1}^\delta(\xi) \leftarrow \sup_{u \in U^\delta} \left\{ \gamma^{\widetilde{\tau}(\xi,u)}. \sum_{\xi'} \widetilde{p}(\xi,\xi',u).V_n^\delta(\xi') + \widetilde{\tau}(\xi,u).\widetilde{r}(\xi,u) \right\}$$

and its difference to the value V^δ of the scheme, is:

$$V^\delta(\xi) - V_{n+1}^\delta(\xi) = \sup_{u \in U^\delta} \left\{ \sum_{\xi'} \left[\gamma^{\tau(\xi,u)} p(\xi,\xi',u).V^\delta(\xi') - \gamma^{\widetilde{\tau}(\xi,u)} \widetilde{p}(\xi,\xi',u).V_n^\delta(\xi') \right] \right.$$
$$\left. + \tau(\xi,u).r(\xi,u) - \widetilde{\tau}(\xi,u).\widetilde{r}(\xi,u) \right\}$$

$$
\begin{aligned}
V^\delta(\xi) - V^\delta_{n+1}(\xi) = \sup_{u \in U^\delta} \Bigg\{ & \gamma^{\tau(\xi,u)} \sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)\right] V^\delta(\xi') \\
& + \left[\gamma^{\tau(\xi,u)} - \gamma^{\widetilde{\tau}(\xi,u)}\right] \sum_{\xi'} \widetilde{p}(\xi,\xi',u).V^\delta(\xi') \\
& + \gamma^{\widetilde{\tau}(\xi,u)} \sum_{\xi'} \widetilde{p}(\xi,\xi',u).\left[V^\delta(\xi') - V^\delta_n(\xi')\right] \\
& + \widetilde{\tau}(\xi,u)\left[r(\xi,u) - \widetilde{r}(\xi,u)\right] + \left[\tau(\xi,u) - \widetilde{\tau}(\xi,u)\right] r(\xi,u) \Bigg\}
\end{aligned}
$$

And from (19), (18) and the Lipschtz property of r, we deduce:

$$
\begin{aligned}
\left|V^\delta(\xi) - V^\delta_{n+1}(\xi)\right| \le \sup_{u \in U^\delta} \Big\{ & \gamma^{\tau(\xi,u)}. \left|\textstyle\sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)\right] V^\delta(\xi')\right| \\
& + \gamma^{\widetilde{\tau}(\xi,u)} \textstyle\sum_{\xi'} \widetilde{p}(\xi,\xi',u).\left|V^\delta(\xi') - V^\delta_n(\xi')\right| \Big\} \\
& + k_\tau \ln \frac{1}{\gamma}.M_{V^\delta}.\delta^2 + \frac{k_2 L_r}{k_1 m_f}\delta^2 + k_\tau M_r \delta^2.
\end{aligned} \tag{21}
$$

Majoration of $\sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)\right] V^\delta(\xi')$:

We have: $V^\delta(\xi') = V^\delta(\xi) + \left[V^\delta(\xi') - V^\delta(\xi)\right]$. But from the properties of the probabilities $p(\xi,\xi',u)$ and $\widetilde{p}(\xi,\xi',u)$, we deduce:

$$
\sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)\right] V^\delta(\xi') = \sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)\right] \left[V^\delta(\xi') - V^\delta(\xi)\right] \tag{22}
$$

Moreover, $|V^\delta(\xi') - V^\delta(\xi)| \le |V^\delta(\xi') - V(\xi')| + |V(\xi') - V(\xi)| + |V(\xi) - V^\delta(\xi)|$. From the theorem 3, the approximation error $\sup_\Omega \left|V^\delta - V\right|$ of the scheme tends to 0 as $\delta \downarrow 0$ for any compact $\Omega \subset O$ and thanks to the continuity of V (theorem 2), $\sup_{\substack{z \in \Omega \\ \|h\| \le \delta}} |V(z) - V(z+h)|$ tends to 0 as $\delta \downarrow 0$.

We deduce: $\left|V^\delta(\xi') - V^\delta(\xi)\right| \le \varepsilon(\delta)$,

with $\varepsilon(\delta) = 2\sup_{z \in \Omega} \left|V^\delta(z) - V(z)\right| + \sup_{\substack{z \in \Omega \\ \|h\| \le \delta}} |V(z) - V(z+h)|$, which tends to 0 as $\delta \downarrow 0$. From (22) and (20), we obtain:

$$
\left|\textstyle\sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)\right] V^\delta(\xi')\right| \le \frac{L_f}{m_f}(k_2 + 2k_N)\delta.\varepsilon(\delta) \tag{23}
$$

The "weak" contraction property (9) and (10) holds :

- Suppose that $\xi \in \Sigma^\delta$: from the property of the exponential function $\gamma^{\Delta t} \le 1 - \frac{\Delta t}{2} \ln \frac{1}{\gamma}$ for Δt small enough, we deduce: $\gamma^{\widetilde{\tau}(\xi,u)} \le 1 - \frac{\widetilde{\tau}(\xi,u)}{2} \ln \frac{1}{\gamma}$, thus $\gamma^{\widetilde{\tau}(\xi,u)} \le 1 - \frac{\delta}{2M_f} \ln \frac{1}{\gamma}$ for small δ, and from (21) and (23) we deduce that:

$$
\left|V^\delta_{n+1}(\xi) - V^\delta(\xi)\right| \le (1 - k.\delta)E^\delta_n + e(\delta).\delta
$$

$$\text{with}: k = \frac{1}{2M_f} \ln \frac{1}{\gamma}$$

$$\text{and}: e(\delta) = \frac{L_f}{m_f}(k_2 + 2k_N)\varepsilon(\delta) + k_\tau \ln \frac{1}{\gamma}.M_{V^\delta}.\delta + \frac{k_2 L_r}{k_1 m_f}\delta + k_\tau M_r \delta$$

Since $\varepsilon(\delta) \downarrow 0$ as $\delta \downarrow 0$, $e(\delta)$ also tends to 0 and the property (9) holds.

– Now suppose that $\xi \in \partial\Sigma^\delta$: from the Lipschitz property of R,

$$\left|V_{n+1}^\delta(\xi) - V^\delta(\xi)\right| = |R(y) - R(\xi)| \leq L_R. \|y - \xi\| \leq L_R.k_N.\delta$$

and the property (10) holds.

Thus the theorem 4 applies and the model-based FD algorithm is convergent. ∎

References

[Bar94] Guy Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques et Applications*. Springer-Verlag, 1994.

[BBS95] Andrew G. Barto, Steven J. Bradtke, and Satinder P. Singh. Learning to act using real-time dynamic programming. *Artificial Intelligence*, (72):81–138, 1995.

[BS91] Guy Barles and P.E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4:271–283, 1991.

[FS93] Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions*. Applications of Mathematics. Springer-Verlag, 1993.

[KD92] Harold J. Kushner and Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*. Applications of Mathematics. Springer-Verlag, 1992.

[MB97] Rémi Munos and Paul Bourgine. Reinforcement learning for continuous stochastic control problems. *Neural Information Processing Systems*, 1997.

[Mun96] Rémi Munos. A convergent reinforcement learning algorithm in the continuous case : the finite-element reinforcement learning. *International Conference on Machine Learning*, 1996.

[Mun97] Rémi Munos. A convergent reinforcement learning algorithm in the continuous case based on a finite difference method. *International Joint Conference on Artificial Intelligence*, 1997.